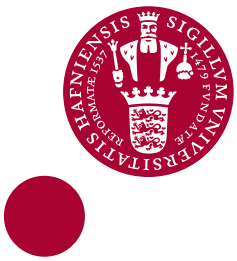

The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies

Andersen, Mikkel Meyer; Eriksen, Poul Svante; Morling, Niels

Published in:
Journal of Theoretical Biology

DOI:
10.1016/j.jtbi.2013.03.009

Publication date:
2013

Journal of Theoretical Biology 329 (2013) 39–51

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies

Mikkel Meyer Andersen [a,*], Poul Svante Eriksen [a,1], Niels Morling [b,2]

[a] Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark
[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, DK-2100 Copenhagen East, Denmark

AUTHOR-HIGHLIGHTS

- The discrete Laplace probability distribution is exploited as an exponential family to make efficient inference.
- The discrete Laplace distribution approximates properties of the Fisher–Wright model of evolution.
- Haplotype frequencies for haploid lineage STR markers are estimated well using a discrete Laplace distribution.
- Open source software to make inference in a mixture of discrete Laplace distributions is supplied.

ARTICLE INFO

Article history:
Received 8 November 2012
Received in revised form
5 March 2013
Accepted 12 March 2013
Available online 21 March 2013

Keywords:
Forensic genetics
Match probability
Likelihood ratio
EM algorithm
Fisher–Wright model

ABSTRACT

Estimating haplotype frequencies is important in e.g. forensic genetics, where the frequencies are needed to calculate the likelihood ratio for the evidential weight of a DNA profile found at a crime scene. Estimation is naturally based on a population model, motivating the investigation of the Fisher–Wright model of evolution for haploid lineage DNA markers. An exponential family (a class of probability distributions that is well understood in probability theory such that inference is easily made by using existing software) called the 'discrete Laplace distribution' is described. We illustrate how well the discrete Laplace distribution approximates a more complicated distribution that arises by investigating the well-known population genetic Fisher–Wright model of evolution by a single-step mutation process. It was shown how the discrete Laplace distribution can be used to estimate haplotype frequencies for haploid lineage DNA markers (such as Y-chromosomal short tandem repeats), which in turn can be used to assess the evidential weight of a DNA profile found at a crime scene. This was done by making inference in a mixture of multivariate, marginally independent, discrete Laplace distributions using the EM algorithm to estimate the probabilities of membership of a set of unobserved subpopulations. The discrete Laplace distribution can be used to estimate haplotype frequencies with lower prediction error than other existing estimators. Furthermore, the calculations could be performed on a normal computer. This method was implemented in the freely available open source software R that is supported on Linux, MacOS and MS Windows.

1. Introduction

The use of haploid lineage DNA markers such as Y-chromosomal short tandem repeats (Y-STRs) or mitochondrial DNA (mtDNA) polymorphisms have important advantages in certain types of forensic genetic casework (Gill et al., 1985; Roewer, 2009; Sibille et al., 2002). If e.g. only a small amount of male DNA is found in combination with a large amount of female DNA, Y-STR typing may be very valuable. If e.g. only a hair shaft is found, mtDNA typing may assist in solving the case. We focus on Y-STR in this paper and note that many of the properties of Y-STR are true for mtDNA as well, because they are both lineage markers.

A very important task in forensic genetics is to evaluate the evidential weight of the evidence by means of likelihood principles (Evett and Weir, 1998; Gill et al., 2001). The likelihood ratio used is

$$LR = \frac{P(E|H_p)}{P(E|H_d)},$$

where H_p is the prosecutor's hypothesis (e.g. 'The suspect is the donor of the genetic data') and H_d is the defense attorney's hypothesis (e.g. 'The suspect is not connected to the crime').

* Corresponding author. Tel.: +45 99408800.
E-mail addresses: mikl@math.aau.dk (M.M. Andersen), svante@math.aau.dk (P.S. Eriksen), niels.morling@sund.ku.dk (N. Morling).
[1] Tel.: +45 99408800.
[2] Tel.: +45 35326115.

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.03.009

In most single doner cases where it is assumed that errors do not happen, it is often assumed that $P(E|H_p)=1$. Then $P(E|H_d)$ is called the 'match probability' and is often interpreted as the probability that an individual drawn randomly from the population has the same DNA profile as the trace found at a crime scene. Note, that if we knew the haplotypes of the entire population, the population frequency of the haplotype in question would be the match probability (in an idealized population without e.g. population structure). Thus, assuming a simple population model, the match probability is the haplotype frequency of the haplotype found at the crime scene.

Due to the lack of recombination, there is statistical dependence between loci, making calculations of match probabilities of lineage markers more challenging than that of autosomal markers (Buckleton et al., 2011). Naïve counts/estimates of match probabilities in a reference database of size n and a haplotype observed x times like x/n, $(x+1)/(n+1)$ or similar seem to be rather conservative and not generally satisfactory (Brenner, 2010; Buckleton et al., 2011). The method of Roewer et al. (2000), Krawczak (2001) and Willuweit et al. (2011) takes the evolutionary aspect of Y-STRs into consideration (see http://www.yhrd.org). Unfortunately, it seems to have some draw backs as indicated by e.g. Andersen (2010). Brenner (2010) suggested a method that takes the rarity of Y-STR haplotypes into consideration. In particular, when considering Y-STR haplotypes comprising a large number of genetic loci, the proportion of haplotypes observed only once – singletons – will be high. Brenner (2010) suggested to adjust/correct the match probability of singletons with a factor, κ, that reflects the ratio between singletons and non-singletons (Robbins, 1968). The κ correction method estimates the match probability by $(1-\kappa)/(n+1)$, where $\kappa=(\alpha+1)/(n+1)$ and α denotes the total number of singletons in the reference database. This method was discussed by Buckleton et al. (2011) and Andersen et al. (2013).

We have developed a model based on assumptions of primarily neutral, single-step mutations of STRs (Ohta and Kimura, 1973) that are following the Fisher–Wright model of evolution (Fisher, 1922, 1930, 1958; Wright, 1931; Ewens, 2004). Caliebe et al. (2010) discussed certain properties of a Fisher–Wright model with neutral single-step mutations. They found the distribution of a quantity that they refer to as the normalized allele process. In this paper, we describe this process and suggest an approximation to its distribution that turn out to be an exponential family called the 'discrete Laplace distribution' due to its similarities to the Laplace distribution of real numbers. This distribution has been described by Inusah and Kozubowski (2006), although they do not note that it is actually an exponential family.

Finally, examples of the use of the discrete Laplace distribution for the estimation of haplotype frequencies for Y-STRs are presented and compared to the results obtained with other methods. The discrete Laplace distribution was used as a family function in a generalized linear model (GLM). The EM algorithm (Dempster et al., 1977) was used to estimate the probability of membership of a set of unobserved subpopulations. The calculations could be performed on a normal computer: Haplotype frequencies of a database with 1000 Y-STR haplotypes consisting of 7 loci could be estimated in around 5 s assuming 1 subpopulation, in around 10 s assuming 2 subpopulations and in around 60 s assuming 5 subpopulations using a Lenovo T410s laptop with 6 GB RAM and an Intel® Core™ i5 CPU model M520 running at 2.40 GHz.

Thus, this paper consists of two parts: (1) an introduction to an exponential family – the discrete Laplace distribution – and (2) an analysis of the application of it in the analyses of lineage markers in population and forensic genetics. Three R (R Development Core Team, 2010) packages 'fwsim' (Andersen and Eriksen, 2012b) (submitted, see Andersen and Eriksen, 2012a for a preprint), 'disclap' (Andersen and Eriksen, 2013a), and 'disclapmix' (Andersen and Eriksen, 2013b) were produced. 'fwsim' (http://cran.r-project.org/package=fwsim) simulates populations under the Fisher–Wright model, 'disclap' (http://cran.r-project.org/package=disclap) implements the exponential family and 'disclapmix' (http://cran.r-project.org/package=disclapmix) uses the EM algorithm (Dempster et al., 1977) to perform inference for a mixture of distributions. Please, refer to Andersen et al. (2013c) for an introduction on how to use these software packages.

2. Discrete Laplace distribution

In this section, the normalized allele process of Caliebe et al. (2010) is described. The discrete Laplace distribution (or double geometric distribution) is introduced as a simple probability distribution. An approximation of the distribution of the normalized allele process in terms of the discrete Laplace distribution is discussed and introduced as an exponential family.

2.1. Motivation

Let N be a constant population size and let $X_g(i)\in\mathbb{Z}$ denote the STR allele (number of repeats) of the ith individual in the gth generation. Thus, it is assumed that alleles are integers. This immediately rules out 'null alleles' (typically a SNP in the primer binding regions of around the Y-STR), intermediate alleles and duplications (Butler, 2005; Budowle et al., 2008). This is a well-known limitation to mathematical STR models that for example coalescent theory also suffers from (Hein et al., 2005; Andersen et al., 2013). The normalized allele process is

$$V_g(i):=X_g(i)-X_g(N) \quad \text{for } i\neq N. \tag{1}$$

The normalized allele process has a mean value of zero. It is a positively recurrent, irreducible, and aperiodic Markov chain that converges exponentially fast to the unique unimodal invariant distribution (Caliebe et al., 2010).

Motivated by the results by Caliebe et al. (2010) – especially the simulation results shown in Caliebe et al. (2010, Figure 1) for certain choices of N, mutation rate, and number of generations—the distribution of the normalized allele process can be approximated by a distribution similar to that of the geometric distribution, but with $\mathbb{Z}$ as support instead of just $\{0,1,\ldots\}$. We refer to this distribution as the 'discrete Laplace distribution'.

2.2. A simple probability distribution

The random variable D follows a discrete Laplace distribution with parameter $0<p<1$ if its probability mass function is such that $P(D=d)\propto p^{|d|}$.

The normalization constant is found by considering the double geometric series

$$\sum_{d\in\mathbb{Z}} p^{|d|}=\frac{1+p}{1-p},$$

such that

$$P(D=d)=\left(\frac{1-p}{1+p}\right)p^{|d|}$$

for $0<p<1$ and $d\in\mathbb{Z}$. Later, in Section 2.5, it is shown that the distribution has mean value

$$\mathbf{E}[D]=\frac{2p}{1-p^2}. \tag{2}$$

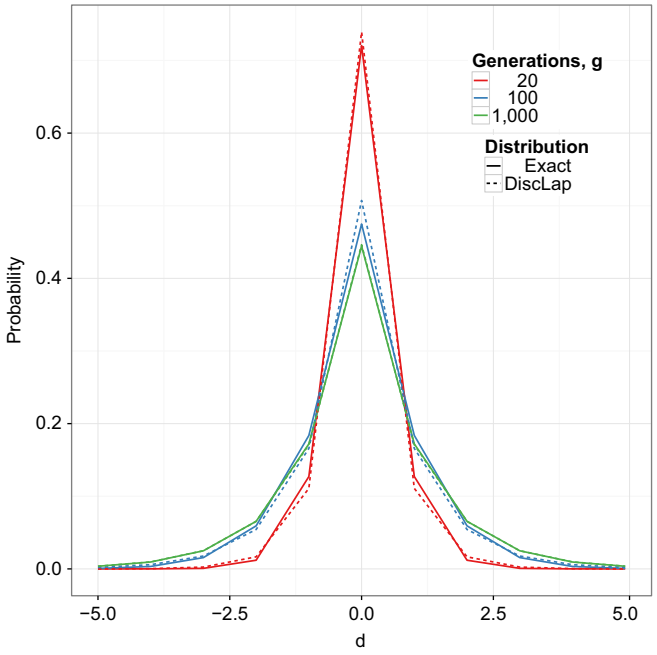

Fig. 1. Exact probability, $\eta_g(d)=P(V_g(i)=d)$, for various values of generations, g, with population size $N=100$ and mutation rate $\mu=0.01$ and the corresponding approximation by the discrete Laplace distribution (DiscLap).

2.3. Approximating the normalized allele process

The interesting quantity is the distribution of Eq. (1), where Caliebe et al. (2010) refer to the probability mass function as η, such that

$$\eta_g(d)=P(V_g(i)=d) \tag{3}$$

for $d\in\mathbb{Z}$. Let $Z_j(i)\in\{-1,0,1\}$ be the mutation event preceding the inheritance of the ith individual in the jth generation. For easier notation, first let

$$Q_j(i)=Z_j(i)-Z_j(N)+2$$

for $d\in\mathbb{Z}$. If μ is the mutation probability, then

$$q(d):=P(Q_j(i)=d)=\begin{cases}\mu^2/4 & \text{if } d=0,\\ \mu-\mu^2 & \text{if } d=1,\\ 1-2\mu+3\mu^2/2 & \text{if } d=2,\\ \mu-\mu^2 & \text{if } d=3,\\ \mu^2/4 & \text{if } d=4,\\ 0 & \text{else.}\end{cases}$$

Thus, $q(d)=r(d-2)$ in the notation of Caliebe et al. (2010) (but as we use r as the number of loci, this function will not be used any further). Let

$$\gamma_g(d)=\eta_g(d+2g). \tag{4}$$

Two expressions of Eqs. (3) and (4) were derived in Caliebe et al. (2010). The first is a recurrence relation Caliebe et al. (2010, Lemma 8). The second is a sum of probability mass function convolutions Caliebe et al. (2010, Theorem 13), which reformulated in terms of γ_g instead of η_g can be expressed as

$$\gamma_g=\frac{1}{N}q*\left(\sum_{i=0}^{g-2}\left[\frac{N-1}{N}\right]^i q^i\right)+\left(\frac{N-1}{N}\right)^{g-1}q^g$$

for $g\in\{2,3,\ldots\}$, where $*$ means the convolution and $q^i=q^{i-1}*q$ means the ith convolution of q.

Using the recurrence relation, Caliebe et al. (2010) plotted this density, which we will compare to an approximation by the discrete Laplace distribution. First, an alternative way of calculating $\eta_g(d)$, and thus $\gamma_g(d)$ numerically, will be described. This method exploits how to do convolutions quickly using a discrete Fourier transformation (Cooley et al., 1969; Brigham, 1988).

By definition

$$\mathbf{E}(\theta^{Q_j})=\sum_{d=0}^{4}P(Q_j=d)\theta^d=\sum_{d=0}^{4}q(d)\theta^d$$

for some $\theta\in\mathbb{C}$, which results in

$$\mathbf{E}(\theta^{\sum_{j=1}^{g}Q_j})=\left(\sum_{d=0}^{4}q(d)\theta^d\right)^g=\sum_{d=0}^{4g}q^g(d)\theta^d$$

due to independence.

Let

$$\theta_a=e^{-2\pi\mathbf{i}a/(4g+1)}$$

for $a=0,1,\ldots,4g$, where $\mathbf{i}$ is the imaginary unit satisfying $\mathbf{i}^2=-1$, and define

$$X_a=\left(\sum_{d=0}^{4}q(d)\theta_a^d\right)^g.$$

Then by Fourier inversion,

$$q^g(d)=\sum_{a=0}^{4g}X_a e^{2\pi\mathbf{i}da/(4g+1)}.$$

Hence, $q^g(d)$ can be found by a fast Fourier transformation (FFT) algorithm, e.g. by using the `fft` function in `R` (R Development Core Team, 2010). When the convolutions are calculated, the value of $\eta_g(d)$ is also quickly calculated.

We suggest that the discrete Laplace distribution approximates the distribution of the normalized allele process, $\eta_g(d)=P(V_g(i)=d)$, in Caliebe et al. (2010). We compared the figures (Caliebe et al., 2010, Figures 1 and 2), see Figs. 1 and 2 with the approximating discrete Laplace distribution. For each set of parameters, the corresponding parameter, p, of the discrete Laplace

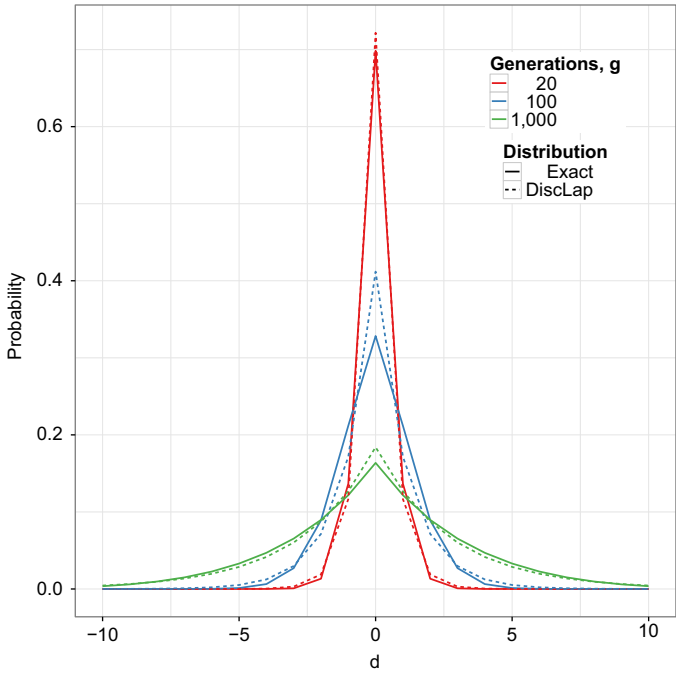

Fig. 2. Exact probability, $\eta_g(d)=P(V_g(i)=d)$, for various values of generations, g, with population size $N=1000$ and mutation rate $\mu=0.01$ and the corresponding approximation by the discrete Laplace distribution (DiscLap).

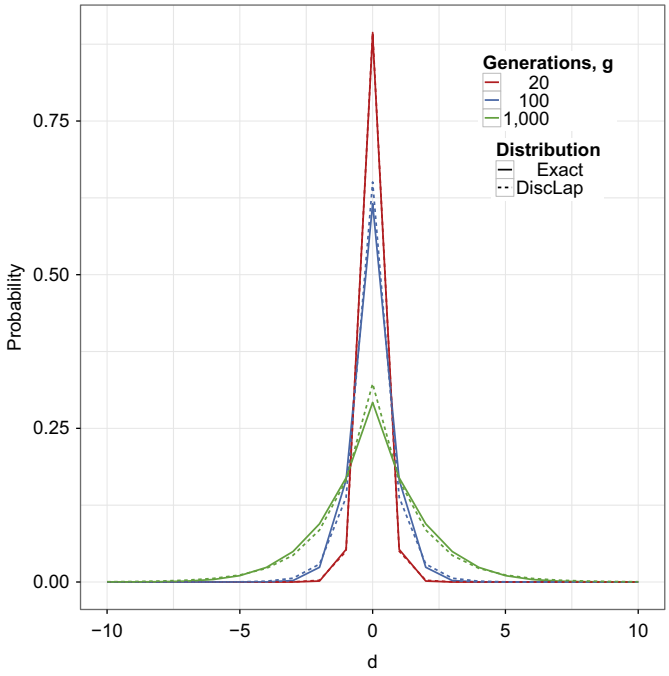

Fig. 3. Exact probability, $\eta_g(d)=P(V_g(i)=d)$, for various values of generations, g, with population size $N=1000$ and mutation rate $\mu=0.003$ and the corresponding approximation by the discrete Laplace distribution (DiscLap).

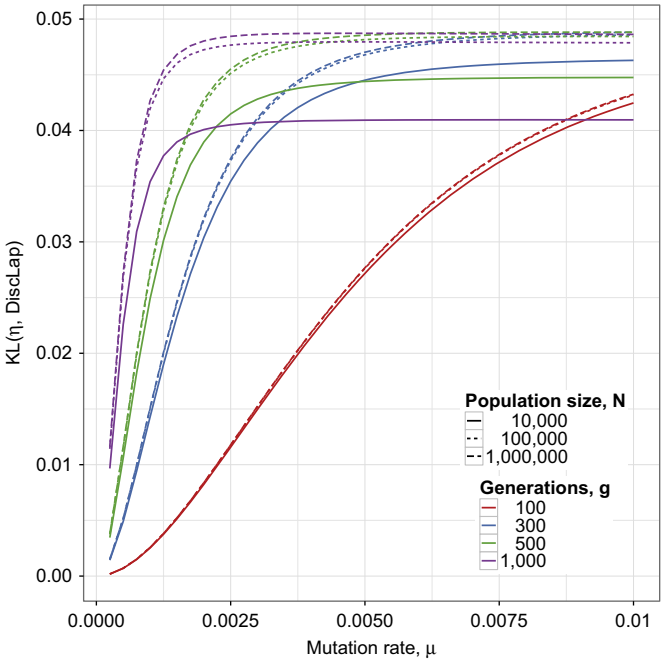

Fig. 4. Kullback–Leibler distance between the exact distribution, η_g, and the approximating discrete Laplace distribution.

distribution was found by calculating the mean

$$\mu=\mathbf{E}[V_g(i)]=2\sum_{d=1}^{2g} d\eta_g(d),$$

and solving Eq. (2) for p to obtain this parameter.

In Fig. 3, a probably more realistic mutation rate for Y-STR of $\mu=0.003$ (Ballantyne et al., 2010) was used.

2.4. Approximation properties

To investigate the approximation properties, the Kullback–Leibler distance (Kullback and Leibler, 1951; Kullback, 1959) between the exact distribution, η_g, given in Eq. (3) (or γ_g given in Eq. (4)) and the discrete Laplace distribution was calculated. Assume that D is distributed according to a discrete Laplace distribution and let $f(d)=P(D=d)$. Let

$$\mathrm{KL}(\eta_g,f)=\sum_{d\in\mathbb{Z}}\eta_g(d)\log\left(\frac{\eta_g(d)}{f(d)}\right)=\sum_{d=-g}^{g}\eta_g(d)\log\left(\frac{\eta_g(d)}{f(d)}\right)$$

as $0\log 0=0$.

The Kullback–Leibler distances for different mutation rates, number of generations and number of individuals are shown in Fig. 4. As seen, the error increases with the mutation rate (to some asymptotic value, it seems). Given a fixed number of generations, the error also increases with the number of individuals. On the other hand, given a fixed number of individuals, there are some points where the lines cross and the number of generations causing the largest error depends on the mutation rate.

2.5. An exponential family

Assume that the signed allele distance, $d\in\mathbb{Z}$, from an ancestor is distributed according to the probability mass function given by

$$f(d;p)=\left(\frac{1-p}{1+p}\right)p^{|d|},\tag{5}$$

where $0<p<1$ is the parameter of the model and $(1-p)/(1+p)$ is the normalizing constant. A reparameterization with

$$\theta=\log p,$$

such that $\theta<0$ shows that this is an exponential family, because

$$f(d;\theta)=\exp\left(\log\left(\frac{1-e^{\theta}}{1+e^{\theta}}\right)+\theta|d|\right)=\exp(\theta|d|-A(\theta))$$

with

$$A(\theta)=\log\left(\frac{1+e^{\theta}}{1-e^{\theta}}\right).$$

The probability mass function `ddisclap`, cumulative distribution function `pdisclap`, random deviates generation function `rdisclap` and family object generation function `DiscreteLaplace` for this exponential family were implemented in the `R` (R Development Core Team, 2010) package `disclap` (Andersen and Eriksen, 2013a).

2.5.1. Cumulants

We now proceed with the cumulants to easily obtain the mean value and the variance function of the distribution. Let D have the probability mass function, $f(d;p)$, as defined in Eq. (5). Then,

$$\mu=\mathbf{E}[D]=\frac{\partial A(\theta)}{\partial\theta}=\frac{\partial p}{\partial\theta}\frac{\partial A}{\partial p}=\frac{2p}{1-p^2}.$$

Furthermore, we obtain the variance function as

$$v(\mu)=\mathbf{Var}[D]=\frac{\partial\mu}{\partial\theta}=\frac{\partial p}{\partial\theta}\frac{\partial\mu}{\partial p}=\mu\left(\frac{1+p^2}{1-p^2}\right).$$

Solving $\mu=2p/(1-p^2)$ for p, yields

$$p=\mu^{-1}\left(\sqrt{\mu^2+1}-1\right),\tag{6}$$

making it possible to obtain the variance function as a function of the mean, i.e.

$$v(\mu)=\mu\sqrt{1+\mu^2}.$$

For practical purposes, in the implementation of the generalized linear model family in R (R Development Core Team, 2010), it is useful to also have the probability mass function as a function of the mean, which is obtained by

$$f(d;p)=\left(\frac{\mu-\sqrt{1+\mu^2}+1}{\mu+\sqrt{1+\mu^2}-1}\right)\times\left(\sqrt{1+\mu^2}-1\right)^{|d|}\mu^{-|d|}.$$

2.5.2. Link function

The canonical link function, g, is found as $g(\mu)=\theta=\log p$, which is equivalent to

$$\theta=g(\mu)=\log\left(\frac{\sqrt{1+\mu^2}-1}{\mu}\right).$$

2.5.3. Deviance

Let

$$L(p;d)=f(d;p)=\left(\frac{1-p}{1+p}\right)p^{|d|}.$$

From Eq. (6),

$$p=p(\mu)=\mu^{-1}\left(\sqrt{\mu^2+1}-1\right),$$

yielding

$$\begin{aligned}l(\mu;d)&=\log L(p(\mu);d)\\&=\log\left(\frac{1-p(\mu)}{1+p(\mu)}\right)+|d|\log(p(\mu))\\&=\log\left(\frac{1-\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)}{1+\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)}\right)\\&\quad+|d|\log\left(\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)\right).\end{aligned}$$

The deviance for one observation, d, is

$$\begin{aligned}D_1(d,p)&=-2\log\frac{L(p(\mu);d)}{L(p(d);d)}\\&=-2(l(\mu;d)-l(d;d))\\&=2(l(d;d)-l(\mu;d)).\end{aligned}$$

In the special case, where $d=0$, we use L'Hôpital's rule (also called Bernoulli's rule) to find the limit using the derivatives of the numerator and denominator and obtain

$$\lim_{d\to0}\frac{\sqrt{d^2+1}-1}{d}=\lim_{d\to0}\frac{d\frac{1}{\sqrt{d^2+1}}}{1}=\lim_{d\to0}\frac{1}{\sqrt{1+\frac{1}{d^2}}}=0$$

such that for $d=0$,

$$l(d;0)=\log 1+0\log 0-\log 1=0$$

and

$$l(\mu;0)=\log\left(\frac{1-\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)}{1+\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)}\right).$$

To summarize

$$D_1(d,p)=\begin{cases}2\log\left(\dfrac{1+\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)}{1-\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)}\right) & \text{if } d=0,\\ 2(l(d;d)-l(\mu;d)) & \text{if } d\neq0.\end{cases}$$

The null deviance for each observation is

$$D_0(d)=\begin{cases}2\log\left(\dfrac{1+\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)}{1-\mu^{-1}\left(\sqrt{\mu^2+1}-1\right)}\right) & \text{if } d=0,\\ 2(l(d;d)-l(\hat{\mu};d)) & \text{if } d\neq0,\end{cases}$$

where $\hat{\mu}$ is the mean of the $|d|$'s.

2.5.4. Parameter estimation

From the theory of exponential families (Azzalini, 1996), for a sample $\{d_i\}_{i=1}^n$ of independent and identically distributed variables following the probability mass function $f(d;p)$ as defined in Eq. (5), the maximum likelihood estimator of $\mu=\mathbf{E}[D]$ is

$$\hat{\mu}=n^{-1}\sum_{i=1}^{n}|d_i|,$$

resulting in the maximum likelihood estimator of p

$$\hat{p}=\hat{\mu}^{-1}\left(\sqrt{\hat{\mu}^2+1}-1\right)$$

by using Eq. (6).

2.5.5. A generalized linear model

With these tools, we can easily define a generalized linear model. This is quite useful, e.g. in (R) (R Development Core Team, 2010), where we can create a family and use the functionality of the (glm) function and its cousins like the prediction function (predict).

3. Estimation of Y-STR haplotype frequencies

In this section, we show how the discrete Laplace family introduced in Section 2.5 can be applied within the field of forensic genetics.

As introduced in Section 2, the normalized allele process $V_g(i)=X_g(i)-X_g(N)$ is the allele difference between any individual n and a fixed individual N. It was empirically validated that the discrete Laplace distribution is an approximation to the distribution of the normalized allele process.

Caliebe et al. (2010) use $X_g(N)$, the allele of the Nth individual, as a reference in the normalized allele process. Note that any other person's allele can be used instead. We choose the reference as the median of all the alleles for one-locus haplotypes (for more loci, it is a bit more complicated and will be treated below). Thus, using the discrete Laplace distribution is merely a qualified guess as the results in Caliebe et al. (2010) will probably not hold when using the median instead of a fixed individual because the median is expected to have lower variance. Below, in Section 3.7, we investigate how qualified the guess actually is.

3.1. Statistical model

Let DL(p,m) be a discrete Laplace model with dispersion parameter $0<p<1$, where we now introduce a location parameter $m\in\mathbb{Z}$. The probability mass function is then

$$f(d;p,m)=\left(\frac{1-p}{1+p}\right)p^{|d-m|}.$$

Inference for a sample, $\{d_i\}_{i=1}^n$, can be made by noticing that the MLE's (maximum likelihood estimates) are

$$\hat{m}=\text{median}\{d_i\}_{i=1}^n,$$

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}|d_i-\hat{m}|\text{ and}$$

$$\hat{p} = \hat{\mu}^{-1}\left(\sqrt{\hat{\mu}^2+1}-1\right),$$

where the equation of $\hat{p}$ stems from Eq. (6).

We will now introduce a model to perform inference in a mixture of multivariate, marginally independent, discrete Laplace distributions.

3.2. Statistical model for multivariate mixtures

Remember that we have r loci instead of just one (mutations across loci are assumed to happen independently). We assume that we have a mixture of c unobserved subpopulations centered at $y_j = (y_{j1}, y_{j2}, \ldots, y_{jr})$ for $j = 1,2,\ldots,c$. We then assume that given a subpopulation, the signed allele distances to the subpopulation center follow independent discrete Laplace distributions.

As before, let $f(d;p)$ be the probability mass function of a DL$(p,0)$ distribution. We define an observation $X = (X_1, X_2, \ldots, X_r)$ to be from a mixture of multivariate, marginally independent, discrete Laplace distributions when the probability of observing $X = x$ is

$$\sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(|x_k - y_{jk}|; p_{jk}),$$

where τ_j is the priori probability for originating from the jth subpopulation. Thus, the parameters of this mixture model are $\{y_j\}_{j=1}^{c}$, $\{\tau_j\}_{j=1}^{c}$ and $\{p_{jk}\}_{\substack{j\in\{1,2,\ldots,c\}\\ k\in\{1,2,\ldots,r\}}}$.

Let MMDL$(c,r,\{y_j\}_{j=1}^{c},\{\tau_j\}_{j=1}^{c},\{p_{jk}\}_{\substack{j\in\{1,2,\ldots,c\}\\ k\in\{1,2,\ldots,r\}}})$ denote such a mixture of multivariate, marginally independent, discrete Laplace distributions.

More theory on finite mixture distributions is given in Titterington et al. (1987).

3.3. Likelihood

In this section, the likelihood of the model is introduced. Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ for $i = 1,2,\ldots,n$ denote the n observed haplotypes from a MMDL$(c,r,\{y_j\}_{j=1}^{c},\{\tau_j\}_{j=1}^{c},\{p_{jk}\}_{\substack{j\in\{1,2,\ldots,c\}\\ k\in\{1,2,\ldots,r\}}})$ distribution. For individual i and subpopulation j, let

$$d_{ijk} = |x_{ik} - y_{jk}|$$

be the distance at the kth locus to the unknown location y_{jk}.

Let z_i denote the (unobserved) subpopulation from which the ith haplotype originated such that $z_i = j$ when the ith haplotype descents from the jth subpopulation. Let

$$v_{ij} = \begin{cases} 1 & \text{if } z_i = j, \\ 0 & \text{otherwise,} \end{cases}$$

such that $v_{i+} = \sum_{j=1}^{c} v_{ij} = 1$.

Let $\tau_j = P(z_i = j)$ denote the a priori probability of originating from the jth subpopulation yielding the constraint $\sum_j \tau_j = 1$. We will soon see that τ_j can be estimated by $\hat{\tau}_j = \hat{v}_{+j}/n = \sum_{i=1}^{n} \hat{v}_{ij}/n$, where $\hat{v}_{ij}$ is an estimate of $P(v_{ij} = 1|x_i)$.

The full likelihood of individual i is given by

$$\begin{aligned} P(x_i, z_i) &= \prod_{j=1}^{c} \left(P(z_i = j) P(x_i | z_i = j)\right)^{v_{ij}} \\ &= \prod_{j=1}^{c} \left(P(z_i = j) \prod_{k=1}^{r} f(d_{ijk}; p_{jk})\right)^{v_{ij}} \\ &= \prod_{j=1}^{c} \tau_j^{v_{ij}} \prod_{k=1}^{r} f(d_{ijk}; p_{jk})^{v_{ij}}, \end{aligned}$$

where $f(d_{ijk}; p_{jk})$ is the probability mass function of the discrete Laplace distribution. Note, that p_{jk} in this case is assumed to depend on locus and subpopulation. We will assume that $\log p_{jk} = \theta_{jk} = \omega_j + \lambda_k$. This means that there is an additive effect of locus and an additive effect of subpopulation and that they do not depend on each other as there is no interaction term. This can be interpreted as ω_j representing the age of the jth subpopulation and λ_k representing the mutation rate at the kth locus.

Hence, the full likelihood of the n independent observations $\{x_i\}_{i=1}^{n}$ is

$$\begin{aligned} L_f &= L_f(\{p_{jk}\}_{j,k}, \{y_j\}_j, \{\tau_j\}_j, \{v_{ij}\}_{i,j}; \{x_i\}_i) \\ &= \prod_{i=1}^{n} P(x_i, z_i) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{c} \tau_j^{v_{ij}} \prod_{k=1}^{r} f(d_{ijk}; p_{jk})^{v_{ij}} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{c} \prod_{k=1}^{r} (\tau_j^{1/r} f(d_{ijk}; p_{jk}))^{v_{ij}}, \end{aligned}$$

where $d_{ijk} = |x_{ik} - y_{jk}|$ and $\log p_{jk} = \omega_j + \lambda_k$.

The marginal likelihood of the observed data is

$$\begin{aligned} L_m &= L_m(\{p_{jk}\}_{j,k}, \{y_j\}_j, \{\tau_j\}_j; \{x_i\}_i) \\ &= \prod_{i=1}^{n} P(x_i) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{c} P(x_i | z_i = j) P(z_i = j) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(d_{ijk}; p_{jk}). \end{aligned} \tag{7}$$

It is a problem that the value of v_{ij} is not known. To deal with this problem, we consider the v_{ij}'s as unobserved variables and use the EM algorithm (Dempster et al., 1977) to estimate the v_{ij}'s.

3.4. Choose subpopulation centers

The simplest way to determine the subpopulation centers, $\{y_j\}_{j=1}^{c}$, is to choose c subpopulation centers and keep these fixed. A more flexible approach is to first choose the initial subpopulation centers, and then allow for the subpopulation centers to be moved around later on if that makes the model better.

Due to the single step mutation model, clustering minimizing the L^1 norm is an obvious choice for initial subpopulation centers as the same mutation rate is assumed for all alleles. This type of clustering is also sometimes referred to as k-medians (the method called k-means is minimizing the L^2 norm). One of the possible methods doing this is the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990), which is supplied by the `R` (R Development Core Team, 2010) library (`cluster`) (Maechler et al., 2005).

A disadvantage of PAM is that the number of subpopulations must be specified beforehand, but one can use BIC (Schwarz, 1978) (Bayesian Information Criteria) to select the best number of subpopulations.

When initial subpopulation centers are chosen, the parameters of the model are estimated using the EM algorithm (Dempster et al., 1977) as described in Section 3.5.

When the EM algorithm has converged, one can try to move the subpopulation centers. Let $\hat{v}_{ij}$ denote the estimate of $P(v_{ij} = 1|x_i)$ after the EM algorithm has converged. Because loci are independent in terms of the mutation process, the total likelihood consists of a product of likelihoods for each locus. This means that we can look at each locus at a time. Let $k \in \{1,2,\ldots,r\}$ be the locus that should be considered.

The MLE of the subpopulation center location assuming all other information is known is then given by

$$\hat{y}_{jk} = \underset{y = \min_i\{x_{ik}\}}{\overset{\max_i\{x_{ik}\}}{\arg\min}} \sum_{i=1}^{n} \sum_{j=1}^{c} \hat{v}_{ij} |x_{ik} - y|,$$

as $g(y) = \sum_{i=1}^{n} \sum_{j=1}^{c} \hat{v}_{ij} |x_{ik} - y|$ is a convex, piecewise linear function that only needs to be evaluated in the ends of each line segment in order to find its minimum.

3.5. EM algorithm

Recall that

$$\mathbf{E}[v_{ij}|x_i] = P(z_i = j|x_i) \quad \text{and} \quad \tau_j = P(z_i = j)$$

and that p depends on locus and subpopulation with no interaction such that $\log p_{jk} = \theta_{jk} = \omega_j + \lambda_k$.

In the following equation, let

$$\mathbf{E}_v := \mathbf{E}_{\{v_{ij}\}_{i,j}|\{x_i\}_i,\{y_j\}_j,\{\tau_j\}_j,\{p_{jk}\}_{j,k}}$$

such that

$$\begin{aligned}\mathbf{E}_v[\log L_f] &= \mathbf{E}_v\left[\log\left(\prod_{i=1}^{n}\prod_{j=1}^{c}\prod_{k=1}^{r}(\tau_j^{\frac{1}{r}} f(d_{ijk};p_{jk}))^{v_{ij}}\right)\right] \\ &= \mathbf{E}_v\left[\sum_{i=1}^{n}\sum_{j=1}^{c} v_{ij}\sum_{k=1}^{r}\log(\tau_j^{\frac{1}{r}} f(d_{ijk};p_{jk}))\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{c}\mathbf{E}[v_{ij}\Big|\{x_i\}_{i=1}^{n}]\times\sum_{k=1}^{r}\log\left(\tau_j^{\frac{1}{r}} f(d_{ijk};p_{jk})\right).\end{aligned}$$

To obtain an estimate of v_{ij}, note that

$$\begin{aligned}\mathbf{E}[v_{ij}|x_i] &= P(z_i = j|x_i) \\ &= \frac{P(z_i = j)P(x_i|z_i = j)}{\sum_{l=1}^{c} P(z_i = l)P(x_i|z_i = l))} \\ &= \frac{\tau_j\prod_k f(d_{ijk};p_{jk})}{\sum_l \tau_l\prod_k f(d_{ilk};p_{lk})},\end{aligned}$$

which gives

$$\hat{v}_{ij} = \frac{\hat{\tau}_j\prod_k f(d_{ijk};\hat{p}_{jk})}{\sum_l \hat{\tau}_l\prod_k f(d_{ilk};\hat{p}_{lk})}$$

by using the estimates $\hat{\tau}_j$ and $\hat{p}_{jk}$ of τ_j and p_{jk}, respectively. For easier notation, let

$$\hat{w}_{ij} = \hat{\tau}_j\prod_k f(d_{ijk};\hat{p}_{jk}) \quad \text{and}$$

$$\hat{v}_{ij} = \frac{\hat{w}_{ij}}{\sum_l \hat{w}_{il}}. \tag{8}$$

And similar to earlier

$$\hat{\tau}_j = \frac{\hat{v}_{+j}}{n}, \tag{9}$$

where $\hat{v}_{+j} = \sum_{i=1}^{n}\hat{v}_{ij}$.

Now, the EM algorithm used can be described:

- *E-step*: Calculate $\hat{v}_{ij}$ using Eq. (8) using the current estimates of $\hat{\tau}_j$ and $\hat{p}_{jk}$ (obtained from the previous E-step and M-step). Now, $\hat{\tau}_j$ can be updated using Eq. (9).
- *M-step*: Maximize

$$L_f = \prod_{i=1}^{n}\prod_{j=1}^{c}\prod_{k=1}^{r}(\tau_j^{1/r} f(d_{ijk};p_{jk}))^{v_{ij}}$$

for $\{p_{jk}\}_{j,k}$ using the current estimates for the other parameters:

$$\begin{aligned}\{\hat{p}_{jk}\}_{j,k} &= \arg\max_{\{p_{jk}\}_{j,k}} L_f \\ &= \arg\max_{\{p_{jk}\}_{j,k}} \prod_{i=1}^{n}\prod_{j=1}^{c}\prod_{k=1}^{r}(\hat{\tau}_j^{1/r} f(d_{ijk};p_{jk}))^{\hat{v}_{ij}} \\ &= \arg\max_{\{p_{jk}\}_{j,k}} \prod_{i=1}^{n}\prod_{j=1}^{c}\prod_{k=1}^{r}(f(d_{ijk};p_{jk}))^{\hat{v}_{ij}}.\end{aligned}$$

This can be done by assuming the GLM model $d_{ijk}\sim\omega_j+\lambda_k$ (other possibilities do exist) with weights $\hat{v}_{ij}$, where $p_{jk} = \exp(\omega_j+\lambda_k)$ (ω_j is a subpopulation effect corresponding to age and λ_k a locus effect corresponding to mutation rate), thus obtaining $\hat{p}_{jk}$.

The assumption that the power $\hat{v}_{ij}$ is equivalent to fixed, known weights in a GLM likelihood is shown in more detail in Wedel and DeSarbo (1995). The `R` (R Development Core Team, 2010) package `FlexMix` (Leisch, 2004; Grün and Leisch, 2008) also uses the same strategy to fit mixtures of GLMs.

According to Dempster et al. (1977, Theorem 1, p. 7), the marginal likelihood Eq. (7) increases with each step of the EM algorithm. Starting values can be chosen as

$$\hat{\tau}_j = 1/c \quad \text{and} \quad \hat{\mu}_{ijk} = d_{ijk}+0.1,$$

where $\hat{\mu}_{ijk}$ is chosen such that the boundary is avoided.

This EM algorithm making inference in a MMDL distribution (mixture of multivariate, marginally independent, discrete Laplace distributions) was implemented in the `R` (R Development Core Team, 2010) package `disclapmix` (Andersen and Eriksen, 2013b).

Note, that there are $cr+(r+c-1)+(c-1)$ parameters in a MMDL distribution: cr for the subpopulation centers

$$\{y_j\}_{j=1}^{c};$$

$(r+c-1)$ for the parameters in the multivariate, marginally independent, discrete Laplace distributions

$$\{p_{jk}\}_{\substack{j\in\{1,2,\ldots,c\}\\ k\in\{1,2,\ldots,r\}}},$$

as there are only main effects of subpopulation and locus; and $c-1$ for the prior probabilities

$$\{\tau_j\}_{j=1}^{c},$$

of originating from each of the c subpopulations, with the reduction of 1 parameter as they sum to 1.

3.6. Haplotype frequency prediction

Given subpopulation centers $\{\hat{y}_j\}_j$, parameters $\{\hat{p}_{jk}\}_{j,k}$ and prior probabilities $\{\hat{\tau}_j\}_j$, e.g. from a converged run of the EM algorithm described in Section 3.5, the haplotype frequency of a haplotype $h = (h_1,h_2,\ldots,h_r)$ with $h_k\in\mathbb{Z}$ for $k\in\{1,2,\ldots,r\}$ can be estimated as

$$\sum_{j=1}^{c}\hat{\tau}_j\prod_{k=1}^{r} f(|h_k-\hat{y}_{jk}|;\hat{p}_{jk}).$$

3.7. Simulation study

To assess the model described in Section 3 for estimating Y-STR (a haploid lineage DNA marker) haplotype frequencies, a simulation study was performed.

A population under the Fisher–Wright model (Fisher, 1922, 1930, 1958; Wright, 1931; Ewens, 2004) with a neutral (in terms of no selection), single step mutation process (Ohta and Kimura, 1973) was simulated using the `R` (R Development Core Team, 2010) package `fwsim` (submitted, see Andersen and Eriksen, 2012a for a preprint). The datasets from this population were sampled and used for estimating haplotype frequencies that were compared to the population frequency.

We simulated 12 different population types by taking all possible combinations of

- Loci: $r=7$
- Mutation rate: $\mu = 0.01$, 0.003 or 0.001
- Generations: $g=500$ or 1000
- Initial population size: $k=10{,}000$ or 50,000.

For all types, the resulting expected population size after g generations was 20,000,000 due to a constant population growth, ρ, that was determined using the number of generations and initial population size as follows. Let N_i denote the population size at the i'th generation. The model from `fwsim` assumes that $N_{i+1}|N_i \sim \text{Poisson}(\rho N_i)$. Thus, if g denotes the number of generations (500 or 1000) and N_0 the initial population size (10,000 or 50,000), then $\mathbf{E}[N_g] = \rho^g N_0$.

For each combination of the parameters, 5 realizations of the population were simulated. For each of these populations, 50 datasets of size 500, 1000 and 5000 were drawn. In total, $12 \cdot 5 \cdot 3 \cdot 50 = 9000$ datasets were sampled and used as basis for comparison.

Note, that the simulated populations are idealized in the sense that the match probability is the haplotype frequency. For all singletons in the dataset, the discrete Laplace distribution approach described in Section 3 was compared to the naïve $1/(n+1)$ estimator and to Brenner's $(1-\kappa)/(n+1)$ estimate, where $\kappa = (\alpha+1)/(n+1)$ and $\alpha+1$ is the number of singletons (haplotypes observed only once) in the dataset as inspired by Robbins (1968). As previously mentioned, the discrete Laplace distribution approach described in Section 3 is implemented in the `R` package `disclapmix` (Andersen and Eriksen, 2013b) that can be used as follows:

```
library (disclapmix)

# Load the dataset simpop
data(simpop)

# The dataset consists of the
# haplotype and the number of
# times it has been observed
head(simpop)

# Make a dataset consisting of
# one observation per row
db <- simpop[
  rep(1:nrow(simpop), simpop$n), 1:7]

# Fit the model with up to 5 subpopulations
res <- disclapmix(db, centers = 1:5,
  use.parallel = TRUE, verbose = 0)

# See the most important information
summary(res$best.fit)

# Predict haplotype frequencies
disclap.estimates <- predict(
  res$best.fit,
  newdata = simpop[, 1:7])
```

For further information on functionality and usage, please run `demo(simpop)` and refer to the documentation `?disclapmix`

As performance measures, the observed bias and the Kullback–Leibler divergence (Kullback and Leibler, 1951; Kullback, 1959) were calculated. Because it is most problematic to estimate the frequency of singletons (haplotypes only observed once), we only focus on these. For a haplotype dataset $H = \{h_i\}_{i=1}^n$ with singletons $\{h_i\}_{i\in S}$ and population frequencies $\{p_i\}_{i\in S}$ estimated as $\{P_{E(H)}(h_i)\}_{i\in S}$ by an estimator E, the bias is

$$B_{H,S}(E) = \frac{1}{|S|}\sum_{i\in S}(P_{E(H)}(h_i) - p_i). \tag{10}$$

The Kullback–Leibler divergence is a measure in information theory about the distance between two probability distributions (we used this distance in Section 2.4) and can also be interpreted as a prediction error. In this case, we only have binary probability distributions. If a haplotype has population frequency p and is estimated to $\hat{p}$, then the Kullback–Leibler divergence is

$$D_{KL}(\hat{p}; p) = \hat{p}\log\left(\frac{\hat{p}}{p}\right) + (1-\hat{p})\log\left(\frac{1-\hat{p}}{1-p}\right).$$

The distribution of Kullback–Leibler divergences for singletons $\{h_i\}_{i\in S}$ is

$$D_{H,S}(E) = \{D_{KL}(P_{E(H)}(h_i); p_i)\}_{i\in S}. \tag{11}$$

The mean and upper 95% quantile of the distribution of Kullback–Leibler divergences for the naïve $1/(n+1)$ estimator, Brenner's κ estimator, and the discrete Laplace based estimator were compared together with the bias.

Note, that the lowest possible prediction error in terms of the Kullback–Leibler divergence is 0, which occurs when $\hat{p} = p$. If this happens for all singletons – that is, all singletons' frequencies were perfectly estimated – then the mean of the Kullback–Leibler divergences would be 0 and so would the bias be. Hence, if the mean of Kullback–Leibler divergence is 0, then so is the bias.

On the other hand, if the bias is 0, then we do not know anything about the Kullback–Leibler divergences. The bias could be 0 if either all the singletons' frequencies were perfectly estimated or if some frequencies were somehow overestimated and others were equally underestimated such that they canceled each other out.

Thus, the prediction error is telling us about the size of the error, whereas the bias is telling us about the direction of the error.

Because migration was not included in the simulation of the populations, only one subpopulation for the discrete Laplace based estimator was used.

3.7.1. Results

As naming convention, DiscLap refers to the model described in Section 3.

For all population types in our simulation study and the performance measures mentioned, the naïve $1/(n+1)$ estimator performed much worse than Brenner's κ estimator and the DiscLap estimator.

Fig. 5 shows estimation in a single dataset (one out of the 9000 datasets analyzed in total). Fig. 6 shows the singleton proportions for the simulated datasets.

The bias as defined in Eq. (10) is shown in Fig. 7. Both the naïve estimator and Brenner's κ estimator seem, in general, to be conservative, which is also what Brenner (2010) states. For dataset size 500, DiscLap seems almost unbiased.

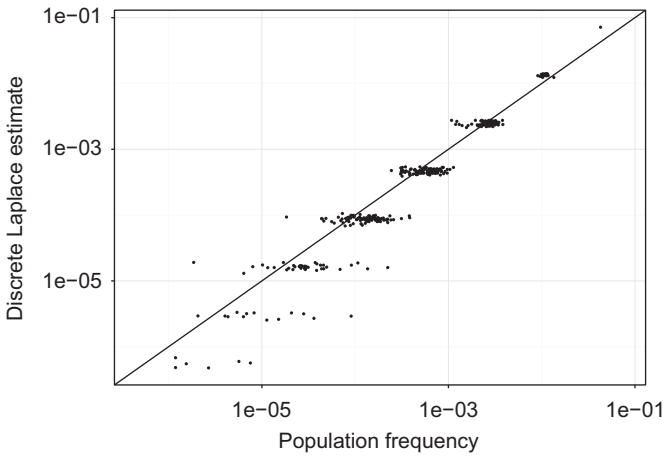

Fig. 5. Haplotype singleton frequency estimation in a single dataset of size 500 from a population with an initial size of 10,000 evolved in 500 generations, a mutation rate of 0.001 and a population growth leading to an expected population size of 20,000,000 after 500 generations. The actual end population size was 19,397,385 consisting of 34,180 different haplotypes.

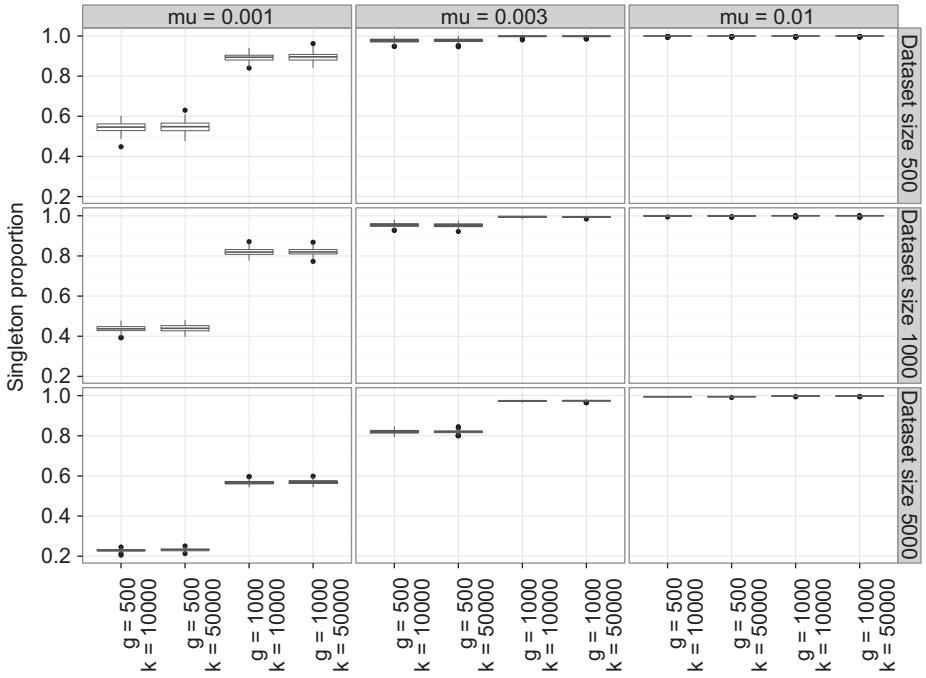

Fig. 6. Singleton proportions of the 9000 simulated datasets.

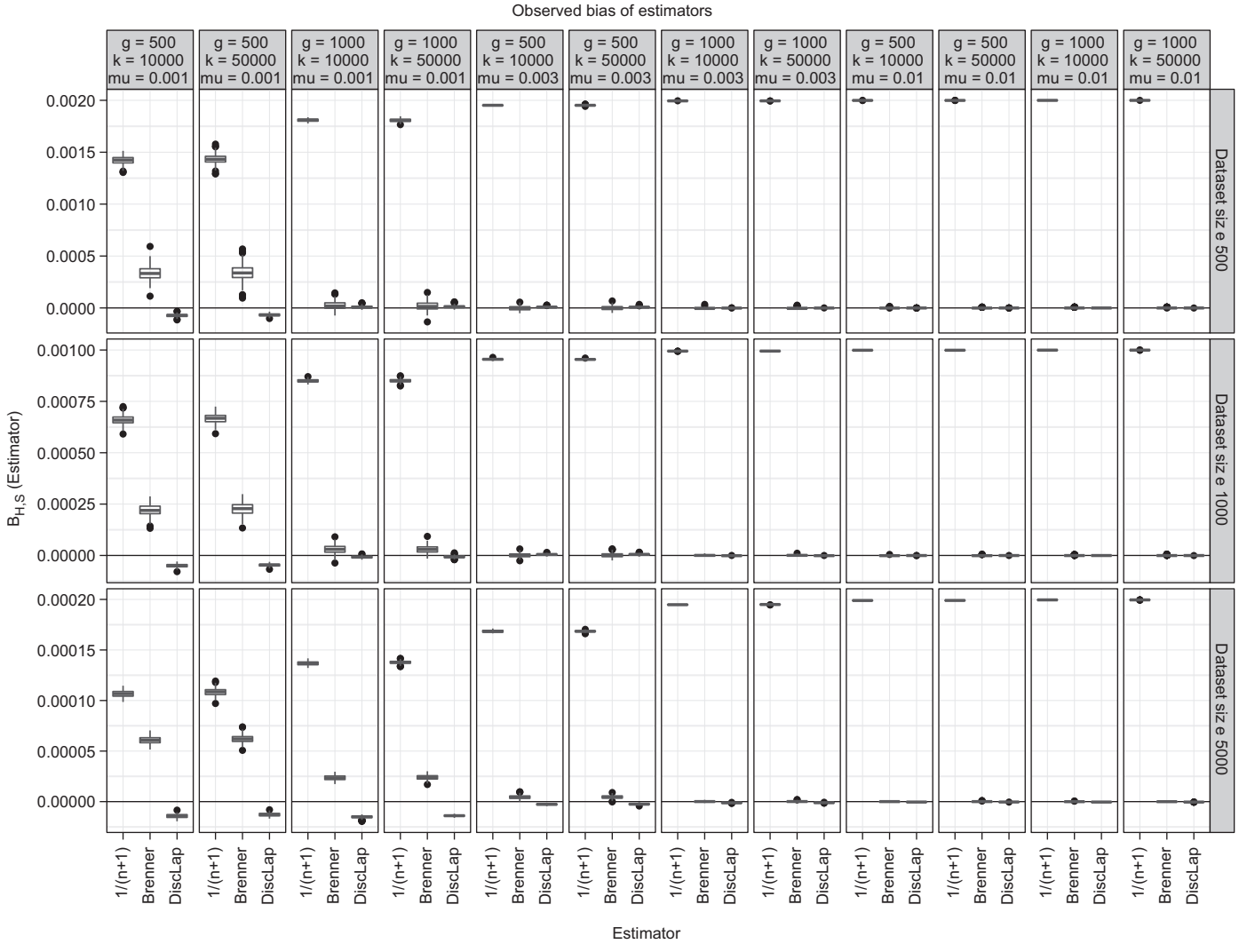

Fig. 7. Bias for the different estimators as defined in Eq. (10).

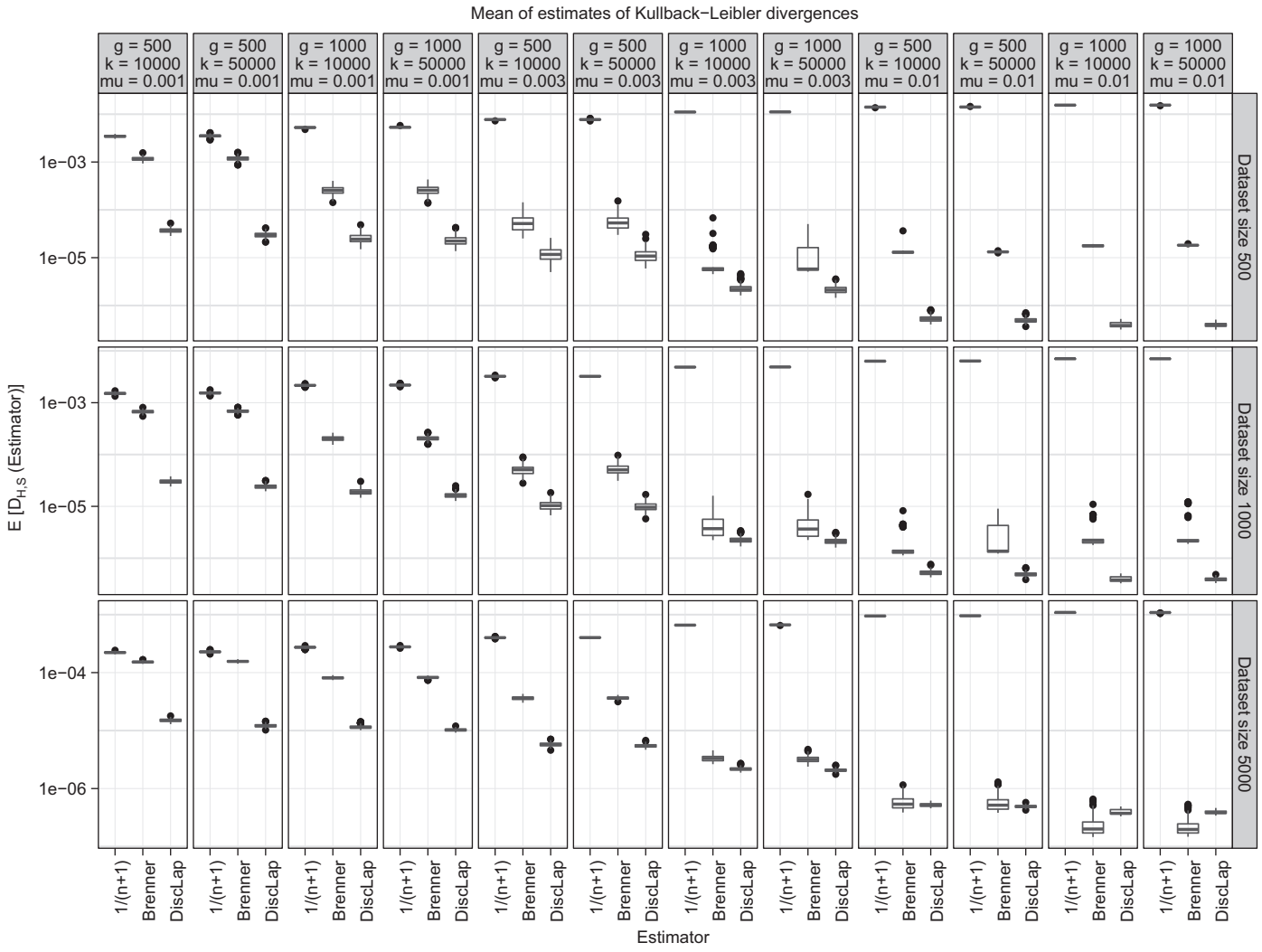

Fig. 8. Mean of the Kullback–Leibler divergences defined in Eq. (11) for each population type. Note, that the ordinate is on a log scale.

This tendency seems stronger for dataset sizes of 1000 and 5000. For the low mutation rate of 0.001, DiscLap seems slightly anti-conservative, whereas for the higher mutation rate of 0.003, it almost seems to be unbiased.

When it comes to the distribution of Kullback–Leibler divergences as defined in Eq. (11), Figs. 8 (the mean) and 9 (the upper 95% quantile) show the same picture, namely that DiscLap overall seems better than Brenner's κ estimator. Table 1 shows a summary of the average proportion between Brenner's κ and DiscLap of the mean of the Kullback–Leibler divergences for each mutation rate and database size.

3.7.2. Discussion

In summary, the prediction error of the estimator using the discrete Laplace distribution (DiscLap) was lower than those of both the κ model by Brenner (2010) and the naïve $1/(n+1)$ estimator. For all population types in our simulation study and the performance measures mentioned (bias and Kullback–Leibler divergence), the naïve $1/(n+1)$ estimator performed much worse than Brenner's κ estimator and the DiscLap estimator.

It seems as if Brenner's κ model estimates haplotype frequencies rather well although it does not incorporate genetic information. One major drawback of this method is that all unobserved haplotypes are assigned the same frequency estimate. Hence, it is doubtful if Brenner's κ model for example is suitable to separate a mixture based on calculating the likelihood ratio (LR) as a measure of the weight of evidence.

Another really important difference between Brenner's κ model and DiscLap is that DiscLap is also able to estimate frequencies for non-singleton haplotypes. Thus, DiscLap can be used no matter if the haplotype has been observed before or not.

In the population types that we studied, we did not observe situations where the estimator based on the discrete Laplace distribution performed worse than the estimator based on Brenner's κ model.

We encourage research on how different population models and migration affects Brenner's κ model and the discrete Laplace distribution.

3.8. Real data example

We analyzed the 1774 German 17-marker haplotypes from release 37 of the YHRD http://www.yhrd.org (Roewer et al., 2001; Willuweit and Roewer, 2009). To render the data usable for both discrete Laplace estimation and the frequency surveying method (Roewer et al., 2000; Krawczak, 2001; Willuweit et al., 2011), some markers and haplotypes were excluded. First, `DYS385a/b` was ignored because of its inherent genotype ambiguity (Roewer et al., 2000) leaving 15 markers for further analysis. Next, 4 haplotypes with 2 alleles reported at `DYS19` and 13 haplotypes with incomplete repeats were excluded, leaving $n=1757$ haplotypes in the data set. Finally, alleles at `DYS389II`

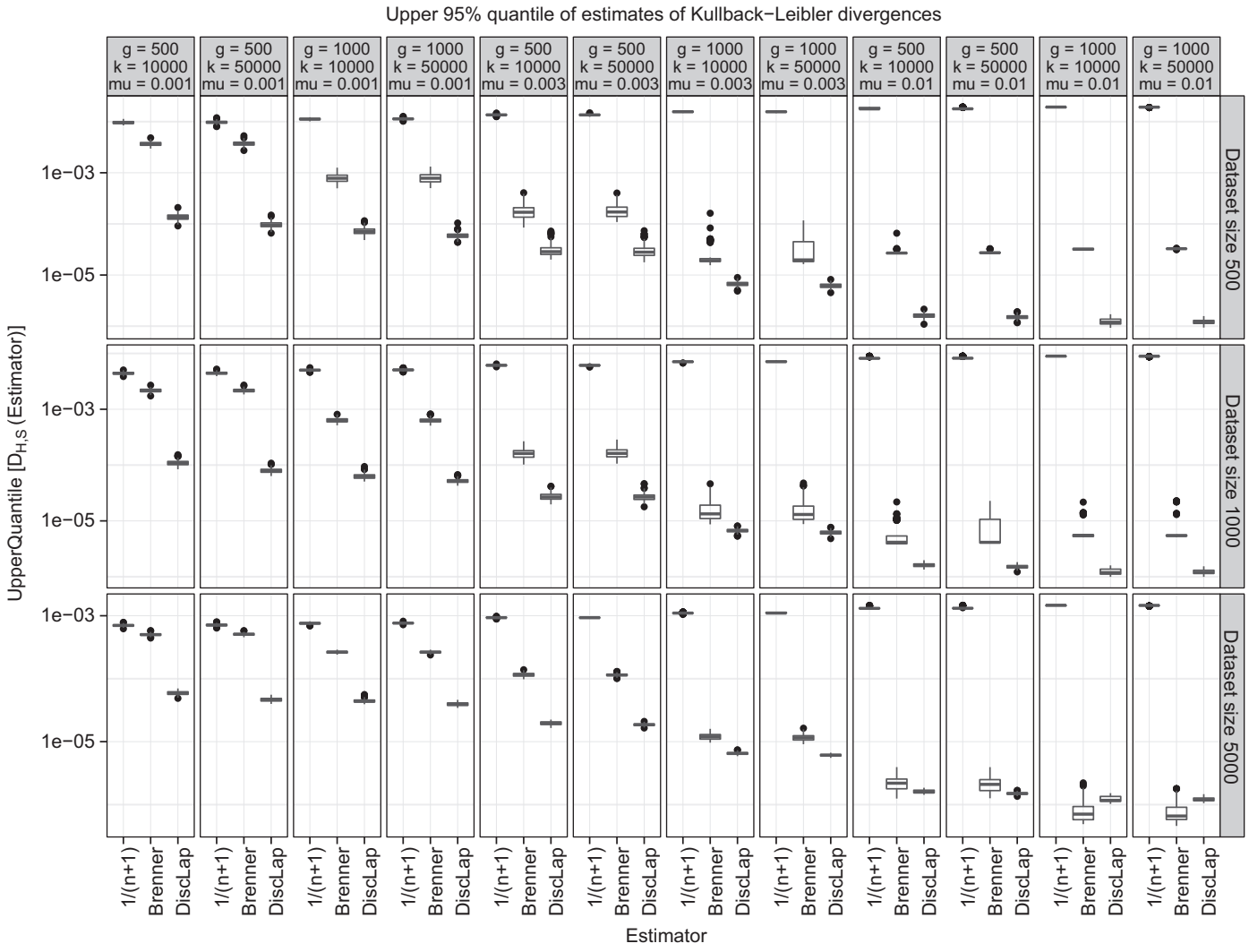

Fig. 9. Upper 95% quantile of the Kullback–Leibler divergences defined in Eq. (11) for each population type. Note, that the ordinate is on a log scale.

Table 1
The average proportion between Brenner's κ and DiscLap of the mean of the Kullback–Leibler divergences for database summarized by mutation rate μ and database size n. A proportion greater than 1 means that the mean of the Kullback–Leibler divergences for Brenner's κ was higher than that of DiscLap. And opposite for a proportion lower than 1.

	$\mu=0.001$	$\mu=0.003$	$\mu=0.01$
$n=500$	23.60	4.88	34.22
$n=1000$	18.67	3.72	5.71
$n=5000$	9.54	4.01	0.86

were replaced by DYS389II minus DYS389I (Butler, 2005). Out of the 1757 haplotypes analyzed, 1469 were singletons.

When restricting the genotype information to the so-called 'minimal haplotype' comprising the seven loci DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393, a total of 392 singletons were observed among the haplotypes of the German data.

3.8.1. Frequency surveying

In its revised form, the surveying method (Willuweit et al., 2011) was based upon an exponential regression model

$$\mu_i = \exp(r_1 W_i + r_2) \quad \text{and}$$
$$\sigma_i = \exp(s_1 W_i + s_2)$$

that links the mean, μ_i, and the standard deviation, σ_i, of the population frequency of the ith haplotype to its weighted inverse molecular distance, W_i, from all other haplotypes in the database. Once the regression parameters, r_1,r_2,s_1,s_2, were determined, the model could serve to define a prior beta distribution of the frequency of any haplotype, h_0, with molecular distance W_0. The parameters of this prior distribution were calculated as

$$\alpha_0 = \frac{\mu_0^2(1-\mu_0)}{\sigma_0^2} - \mu_0 \quad \text{and}$$

$$\beta_0 = \alpha_0\left(\frac{1-\mu_0}{\mu_0}\right).$$

Maximum likelihood estimates of the regression parameters were obtained in our study by numerical optimization (Willuweit et al., 2011) using the Nelder–Mead simplex algorithm with up to 1500 iterations as implemented in R (R Development Core Team, 2010). Several different starting values of (r_1,r_2,s_1,s_2) were tried, and the vector resulting in the highest likelihood was chosen. The starting values were taken from the Cartesian product $\{15,20,30.82\}\times\{-10,-15,-13.17\}\times\{15,20,28.95\}\times\{-10,-15,-11.71\}$, where the last elements in the sets are the respective binning estimates of the Western European population given in Table 3 of Willuweit et al. (2011).

Let n_i be the number of times that the ith haplotype was observed in the database with $n=\sum_i n_i$ being equal to the database size. For comparison with the other estimators, we used the mean of the posterior Beta$(\alpha_i+n_i-1,\beta_i+n-n_i)$ given by

$$\frac{\alpha_i+n_i-1}{\alpha_i-1+\beta_i+n}$$

as the haplotype surveying estimate of the population frequency of haplotype, h_i.

3.8.2. Results

For both the full and the minimal haplotype, only the singletons were used to compare the haplotype frequency estimates provided by the different estimators.

Fig. 10 shows the results of the 7-loci-database. Fig. 11 shows the results of the 15-loci-database. It is impossible to make any sensible conclusion from this as we do not know the true haplotype frequencies.

4. Discussion

The first part of this paper describes an exponential family called the discrete Laplace distribution. The fact that the discrete Laplace distribution is an exponential family makes inference somewhat easier as theory on exponential families already exists and can be exploited. This also means simpler and faster computer software because existing implementations that have been optimized can be used.

The second part of this paper consists of an application of the discrete Laplace distribution, namely how to estimate Y-STR haplotype frequencies. An estimate of the frequency of a Y-STR haplotype can be used as an estimate of the match probability (assuming an idealized population without population substructure), which is an essential part in forensic genetics when evaluating the evidential weight of the evidence by means of likelihood principles. The calculations could be performed on a normal computer. We demonstrate that for our simulation study on 12 different population types (varying mutation rate, population growth and generations) resulting in 9000 datasets (of size 500, 1000 and 1500), the haplotype frequency estimation based on the discrete Laplace distribution performs overall better than the κ model by Brenner (2010). The mean of the Kullback–Leibler divergences is in general lower for the estimation based on the discrete Laplace distribution than that based on Brenner's κ cf. Table 1.

Furthermore and very importantly, Brenner's κ can only be used for singletons whereas estimation based on the discrete Laplace distribution can be used for all haplotypes.

We encourage research on how different population models and migration affects Brenner's κ model and the discrete Laplace distribution.

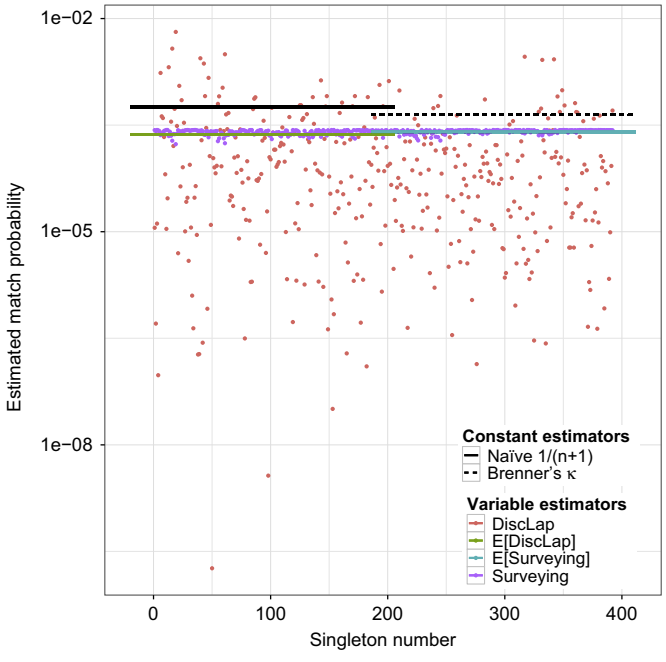

Fig. 10. Comparison of the haplotype frequency estimators for the 7 loci German database consisting of 1757 haplotypes of which 392 were singletons. Thus, Brenner's $\kappa=4.4\times10^{-4}$. Note, that the ordinate with the estimated haplotype frequency is on a log scale. 11 subpopulations were used (1 through 15 subpopulations were tried, 11 subpopulations had the lowest BIC score Schwarz, 1978). The line 'E[DiscLap]' refers to the average of the DiscLap estimates and the line 'E[Surveying]' refers to the average of the surveying estimates.

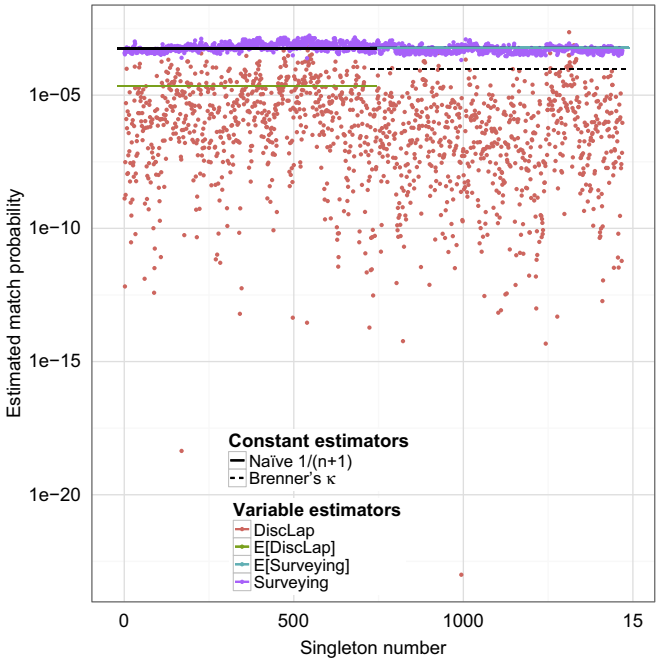

Fig. 11. Comparison of the haplotype frequency estimators for the 15 loci German database consisting of 1757 haplotypes of which 1469 were singletons. Thus, Brenner's $\kappa=9.3\times10^{-5}$. Note, that the ordinate with the estimated haplotype frequency is on a log scale. 14 subpopulations were used (1 through 15 subpopulations were tried, 14 subpopulations had the lowest BIC score Schwarz, 1978). The line 'E[DiscLap]' refers to the average of the DiscLap estimates and the line 'E[Surveying]' refers to the average of the surveying estimates.

Acknowledgments

We thank Amke Caliebe, Christian-Albrechts-Universität zu Kiel, Germany, for providing the R code needed to estimate haplotype frequencies using the surveying approach.

References

Andersen, M.M., 2010. Y-STR: Haplotype Frequency Estimation and Evidence Calculation. Master's Thesis. Aalborg University, Denmark.

Andersen, M.M., Caliebe, A., Jochens, A., Willuweit, S., Krawczak, M., 2013. Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. Forensic Sci. Int.: Genet. 7, 264–271.

Andersen, M.M., Eriksen, P.S., 2012a. Efficient forward simulation of Fisher–Wright populations with stochastic population size and neutral single step mutations in haplotypes. Preprint. ArXiv:1210.1773.
Andersen, M.M., Eriksen, P.S., 2012b. FWSIM: Fisher–Wright population simulation. R package version 0.2-5.
Andersen, M.M., Eriksen, P.S., 2013a. disclap: Discrete Laplace Family. R package version 1.2.
Andersen, M.M., Eriksen, P.S., 2013b. disclapmix: discrete Laplace mixture inference using the EM algorithm. R package version 0.3.
Andersen, M.M., Eriksen, P.S., Morling, N., 2013c. A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies. Preprint. arXiv:1304.2129.
Azzalini, A., 1996. Statistical Inference—Based on the Likelihood. Chapman & Hall.
Ballantyne, K.N., et al., 2010. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am. J. Hum. Genet. 87, 341–353.
Brenner, C.H., 2010. Fundamental problem of forensic mathematics – the evidential value of a rare haplotype. Forensic Sci. Int. Genet. 4, 281–291.
Brigham, E.O., 1988. The Fast Fourier Transform and its Applications. Prentice Hall.
Buckleton, J., Krawczak, M., Weir, B., 2011. The interpretation of lineage markers in forensic DNA testing. Forensic Sci. Int.: Genet. 5, 78–83.
Budowle, B., Aranda, X., et al., 2008. Null allele sequence structure at the DYS448 locus and implications for profile interpretation. Int. J. Legal Med. 122, 421–427.
Butler, J.M., 2005. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, 2nd ed Academic Press.
Caliebe, A., Jochens, A., Krawczak, M., Rösler, U., 2010. A Markov chain description of the stepwise mutation model: local and global behaviour of the allele process. J. Theoretical Biol. 266, 336–342.
Cooley, J., Lewis, P., Welch, P., 1969. The finite Fourier transform. IEEE Trans. Audio Electroacoust. 17, 77–85.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. 39, 1–38.
Evett, I.W., Weir, B.S., 1998. Interpreting DNA Evidence. Sinauer Associates.
Ewens, W.J., 2004. Mathematical Population Genetics. Springer.
Fisher, R.A., 1922. On the dominance ratio. Proc. R. Soc. Edin. 42, 321–341.
Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.
Fisher, R.A., 1958. The Genetical Theory of Natural Selection, 2nd revised ed. Dover, New York.
Gill, P., Brenner, C., Brinkmann, B., Budowle, B., Carracedo, A., Jobling, M., de Knijff, P., Kayser, M., Krawczak, M., Mayr, W., Morling, N., Olaisen, B., Pascali, V., Prinz, M., Roewer, L., Schneider, P., Sajantila, A., Tyler-Smith, C., 2001. DNA commission of the international society of forensic genetics: recommendations on forensic analysis using Y-chromosome STRs. Forensic Sci. Int. 124, 5–10.
Gill, P., Jeffreys, A.J., Werrett, D.J., 1985. Forensic application of DNA fingerprints. Nature 318, 577–579.
Grün, B., Leisch, F., 2008. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Software, 28.
Hein, J., Schierup, M.H., Wiuf, C., 2005. Gene Genealogies Variation, and Evolution: A Primer in Coalescent Theory. Oxford University Press.
Inusah, S., Kozubowski, T.J., 2006. A discrete analogue of the Laplace distribution. J. Stat. Plann. Inference 136, 1090–1102.
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.
Krawczak, M., 2001. Forensic evaluation of Y-STR haplotype matches: a comment. Forensic Sci. Int. 118, 114–115.
Kullback, S., 1959. Information Theory and Statistics. John Wiley and Sons.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.
Leisch, F., 2004. FlexMix: a general framework for finite mixture models and latent class regression in R. J. Stat. Software, 11.
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., 2005. Cluster analysis basics and extensions. Rousseeuw et al. provided the S original which has been ported to R by Kurt Hornik and has since been enhanced by Martin Maechler: speed improvements, silhouette() functionality, bug fixes, etc. See the 'Changelog' file (in the package source).
Ohta, T., Kimura, M., 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22, 201–204.
R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Robbins, H.E., 1968. Estimating the total probability of the unobserved outcomes of an experiment. Ann. Math. Stat. 39, 256–257.
Roewer, L., 2009. Y chromosome STR typing in crime casework. Forensic Sci. Med. Pathol. 5, 77–84.
Roewer, L., Kayser, M., de Knijff, P., et al., 2000. A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. Forensic Sci. Int. 114, 31–43.
Roewer, L., Krawczak, M., Willuweit, S., et al., 2001. Online reference database of European Y-chromosomal short tandem repeat STR haplotypes. Forensic Sci. Int. 2–3, 106–113.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Sibille, I., Duverneuil, C., et al., 2002. Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. Forensic Sci. Int. 125, 212–216.
Titterington, D.M., Smith, A.F.M., Makov, U.E., 1987. Statistical Analysis of Finite Mixture Distributions. Wiley.
Wedel, M., DeSarbo, W.S., 1995. A mixture likelihood approach for generalized linear models. J. Classification 12, 21–55.
Willuweit, S., Caliebe, A., Andersen, M.M., Roewer, L., 2011. Y-STR frequency surveying method: a critical reappraisal. Forensic Sci. Int.: Genet. 5, 84–90.
Willuweit, S., Roewer, L., 2009. Y chromosome haplotype reference database (YHRD): update. Forensic Sci. Int.: Genet. 1, 83–87.
Wright, S., 1931. Evolution in mendelian populations. Genetics 16, 97–159.